Multi-partner Demonstration of BGP-LS enabled multi-domain EON control and instantiation with H-PCE

O. González de Dios, R. Casellas, R.Morro, F. Paolucci, V. López, R. Martínez, R. Muñoz, R. Vilalta, P. Castoldi

***Abstract*— The control of Multi-domain Elastic Optical Networks (EON) is possible by combining H-PCE based computation, BGP-LS topology discovery, remote Instantiation via PCEP, and signaling via RSVP-TE. Two evolutionary architectures are considered, one based in stateless H-PCE, PCEP Instantiation and end-to-end RSVP-TE signaling (SL-E2E), and a second one based on stateful active H-PCE with per-domain instantiation and stitching.**

This paper presents the first multi-platform demonstration that fully validates both control architecture achieving multi-protocol interoperability. SL-E2E leads to slightly faster provisioning, but needs to keep the state of the stitching of the e2e LSPs in the parent PCE.

Index Terms—Multi-domain EON, H-PCE, BGP-LS, PCEP

I. Introduction

Elastic Optical Networks (EON) represent the state-of-the-art of connection-oriented optical networks. Thanks to the recent advances in the design of flexible bandwidth-variable transponders (BVT), capable of transmitting/receiving signals with configurable physical parameters (i.e., bitrate, modulation format), and to the availability of Spectrum Selective Switches (SSS), capable of switching frequency slices multiple of 12.5GHz, EON have the potential to enable fully configurable multi-bitrate lightpaths thus increasing service-oriented flexibility and overall network capacity [1].

Adequate control plane operation is required to achieve lightpath provisioning, along with additional EON-specific procedures, such as elastic operation and re-optimization (e.g., defragmentation). In the context of the ABNO architecture [3], a centralized multi-component controller is envisioned and the Path Computation Element (PCE) may be used, besides path computation, as functional component for direct lightpath instantiation and release [5]. In the case of multi-domain networks, hierarchical approach is considered for path computation. Hierarchical PCE (HPCE), comprising a parent PCE (pPCE) and several per-domain Child PCEs (cPCE), has been successfully demonstrated for WSON [6]. However, both optimal domain sequence and intra-domain segment selection could be achieved only if pPCE is aware of detailed Traffic Engineering (TE) topology. Such information may be available at pPCE without scalability issues by resorting to the recently proposed TE Link State Information extensions to BGP (BGP-LS) [7], as experimentally demonstrated in [8].

In this paper, two extended HPCE architectures are considered to control multi-domain EON based on GMPLS. The first architecture (StateLess-H-PCE with e2e signaling and instantiation, shortened as SL-E2E), is based on stateless PCEs and PCEP instantiation that triggers end-to-end RSVP-TE signaling. The second one (StateFul-H-PCE with Per-Domain instantiation and stitching, shortened as SF-PD), considers stateful PCEs with active capabilities, and a per-domain instantiation, with intra-domain RSVP-TE signaling.

Routing (OSPF-TE and BGP-LS), path computation/instantiation (PCEP) and signaling (RSVP-TE) protocol extensions necessary to enable inter-domain EON lightpath provisioning are detailed. The two proposed choices for provisioning are compared and experimentally validated.

For the first time, such extensions are evaluated in a complete distributed multi-partner control plane test-bed in order to fully validate inter-operability among multi-platform and/or multi-vendor network/node controllers. Complete provisioning performances are detailed, including BGP-LS topology update, path computation (i.e., segment computation, selection and path concatenation, including BVT end-point indication and configuration), instantiation, end-to-end RSVP-TE signaling and per-domain instantiation and stitching.

O. González de Dios and V. López are with Telefónica Research and Development (I+D), in the Planning of IP & Transport Networks department of Telefonica Global CTO, Edificio Sur, Distrito Telefonica, 28050 Madrid (Spain).

R. Casellas, R. Martínez, R. Muñoz, R. Vilalta, are with CTTC. Av. Carl Friedrich Gauss n7, 08860 Castelldefels, Barcelona.

R. Morro is with Telecom Italia, Via G. Reiss Romoli 274, 10148 Torino, Itay

F. Paolucci and P. Castoldi are with CNIT-Scuola Superiore Sant'Anna, Pisa, Italy.

A. Hierarchical PCE for Multi-Domain

The H-PCE has been retained as the most suitable technology to compute optimum routes for LSPs crossing multiple domains. The pPCE is responsible for domain sequence computation. Then, in each identified domain, a child PCE (cPCE) performs segment expansion. The pPCE exploits an abstracted domain topology map that contains the child domains and their interconnections. Several innovative enhancements to this approach are under investigation.

o First, besides reachability information, a mesh of abstracted links between border nodes is introduced in the parent TED to improve the effectiveness of domain sequence computation.

o Second, the north-bound distribution of Link-State and TE information using BGP (i.e., BGP-LS) is the protocol solution proposed to provide link information to the pPCE.

o Third, specific extensions to BGP-LS for elastic optical networks are also introduced.

B. Stateless vs Stateful/Active PCE

The PCE architecture was proposed to provide effective constraint-based path computations. So far, the PCE has been mainly deployed with a stateless architecture, i.e. the PCE only relies on the TED which includes information on resource utilization. More recently, the PCE architecture has been extended with stateful capabilities, enabling the attributes of the established LSPs (e.g., the route) to be stored and maintained at the LSP State Database (LSPDB) [6]. Furthermore, a stateful PCE may also include the active functionality which enables the PCE to issue recommendations to the network, e.g. to dynamically update LSP parameters through the PCE Communication Protocol (PCEP). In the IDEALIST project [2], the (active) stateful architecture has been adopted to enable a number of advanced traffic engineering functionalities, including elastic LSP operations and global defragmentation in flexi-grid networks. For example, the PCE is able to account for the actual network conditions, run complex re-optimization algorithms, and operate on existing LSPs to reduce the overall network fragmentation. The implementation of the stateful functionality has also to account for some deployment considerations, mainly related to reliability, synchronization (e.g., after restart) and scalability issues. In terms of scalability, the stateful PCE is not designed to be operated over the entire Internet. On the contrary, its domain of visibility has to be adequately dimensioned, considering a sufficiently over-provisioned system.

C. Multi-domain Provisioning

Inter-domain TE LSPs can be supported by one of three options: contiguous LSPs, stitched LSPs and nested LSPs. In the flex-grid context, the latter solution is not applicable. Since these solutions require a high degree of control plane interoperability both for routing and for signaling, we are considering:

1) Instantiation + RSVP-TE end-to-end

This approach employs a single RSVP-TE end-to-end session for setting up the connection between the domains. The H-PCE framework is used for path computation, even if in a stateful mode, in which local end multi-domain LSPDBs are maintained, and PCEP for LSP provisioning demanding the Path establishment to the ingress node. This allows simplified setup and teardown procedures, especially in case of exceptions handling, w.r.t the case of separated signaling sessions (one per domain) at the cost of the need for interoperability also at RSVP-TE level.

2) Per-domain instantiation + stitching

This approach takes full advantage of the H-PCE framework in which the pPCE orchestrates the involved cPCEs, acting as the responsible within their own domain, for the establishment (and release) of connections by means of an underlying GMPLS control plane. In this case, all PCEs are stateful and must have instantiation capabilities and every domain has its own "local" RSVP-TE session. The data plane connectivity is insured by the concatenation of media channels at each domain, while the coordination among the domains (i.e. ingress/egress ports, labels, etc.) is under the responsibility of the pPCE. In this case, interoperability requirements are scoped to PCEP extensions for stateful PCE with instantiation capabilities and no protocols are required at the inter-domain boundaries.

II. Architecture and Control Plane Procedures

Two architectures, both built on the H-PCE concept, are considered to solve the problem of multi-domain control. The first one considers a stateless PCE and end-to-end signaling from the head-end node. The second one is an evolution of the previous, and considers a stateful active parent PCE and a per-domain provisioning through remote instantiation. The parent PCE is the control entity that will have the notion of the multi-domain LSP and each domain will be aware only of its segment. The architecture and control plane procedures are described below.

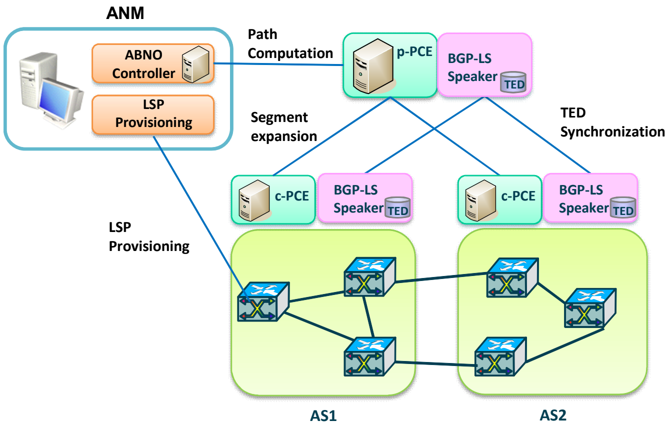


Figure 1. SL-E2E Control Architecture

A. Stateless H-PCE with end-to-end signaling and instantiation

The first proposed architecture, Stateless-H-PCE with e2e signaling and instantiation, (shortened as SL-E2E) is built around the hierarchical PCE (HPCE) framework, in which a parent PCE (pPCE) coordinates several children PCEs (cPCE), one per network domain (Figure 1). The pPCE is in charge of domain selection and inter-domain path computation. Children PCEs are responsible for segment expansion, i.e. for path computation in their respective domains. BGP-LS is deployed for topology

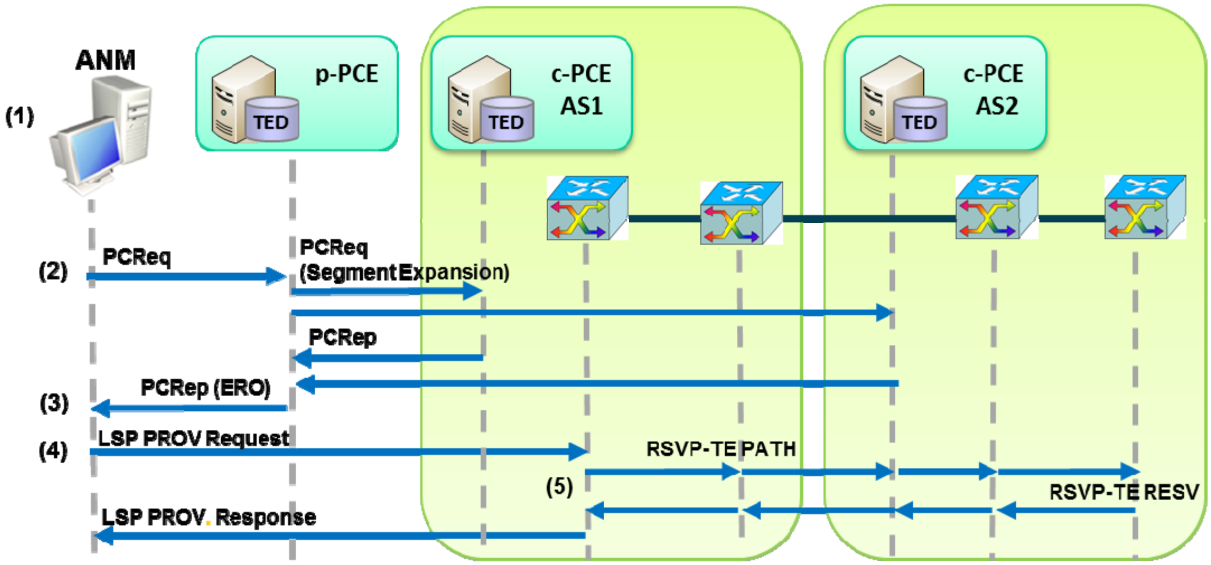


Figure 2. Message flow in Stateless H-PCE with end-to-end signaling and instantiation (SL-E2E) architecture

abstraction and to export inter-domain TE information to the pPCE. LSP provisioning is triggered by an SDN controller that monitors network resources utilizations and is able to decide the optimal network configuration based on the status, bandwidth availability and user service. It leverages recently proposed extensions to the PCEP protocol for the so-called stateful and active PCE and includes instantiation capabilities. A single end-to-end RSVP-TE signaling session is used for setting up the connection between the domains; this allows simplified setup and teardown procedures, especially in case of exceptions handling, w.r.t the case of separated signaling sessions (one per domain).

1) SL-E2E Procedure

Let us detail the main procedure for the establishment of an LSP with the help of Figure 2, as implemented in the multi-partner test-bed. Upon request (1), the SDN controller (referred as Adaptive Network Manager, ANM in the figure) triggers the provisioning. First, it requests a multi-domain path computation (2), which is a two-step process in which the pPCE obtains the domain sequence and then requests the children to expand the domain path within their respective domains. The pPCE composes and end-to-end ERO, which, by default, uses unnumbered interfaces represent outgoing TE links but, to convey information about the ingress port, it can be prepended with an additional ingress interface (facing the client) and may either end with a IPv4 prefix address or an unnumbered interface meaning that the LSP ends at the output interface with an additional cross-connect. Explicit label control (ELC) conveys information about the outgoing label (frequency slot) that will be used by the downstream node in switching. The actual LSP provisioning takes place after the end to end path has been computed (3), and the provisioning manager uses the PCEP interface with the ingress node to request a Path establishment (4). It is based on the use of PCInitiate and the PCRpt messages: the PCInitiate includes the SRP, LSP, ENDPOINTS, ERO objects, and instructs the ingress node to initiate the signaling procedure, based on the Path/Resv RSVP-TE message exchange with an end-to-end session (5). Upon completion of the signaling process, the PCRpt message is sent back to the provisioning manager, additionally including the route object RRO and the allocated frequency slot.

B. StateFul-H-PCE with Per-Domain instantiation and stitching

In the second architecture, StateFul-H-PCE with Per-Domain instantiation and stitching (SF-PD), a stateful condition is introduced at the pPCE to enable advanced TE solutions, e.g. multi-domain re-optimization. The approach completes the hierarchical path computation composed of domain sequence selection and segment expansion with a subsequent route segmentation and segment provisioning, as explained next.

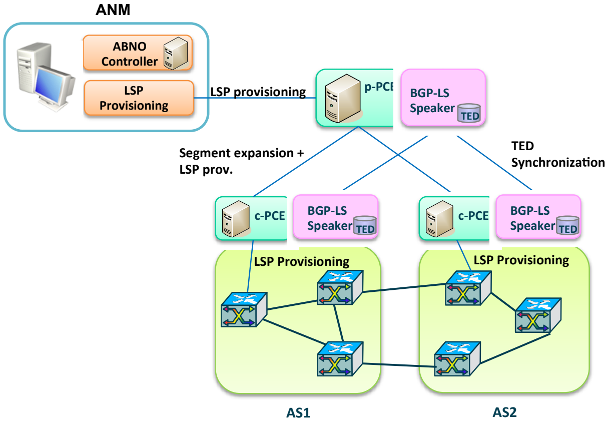


Figure 3. SF-PD Control Architecture

The multi-domain LSP provisioning is based on instantiation extensions to PCEP. A single connection happens by stitching "on the wire" as many segments as required. The systems thus proceed in a two-step process, a full path computation (the strict ERO is retrieved by p-PCE) and the subsequent provisioning.

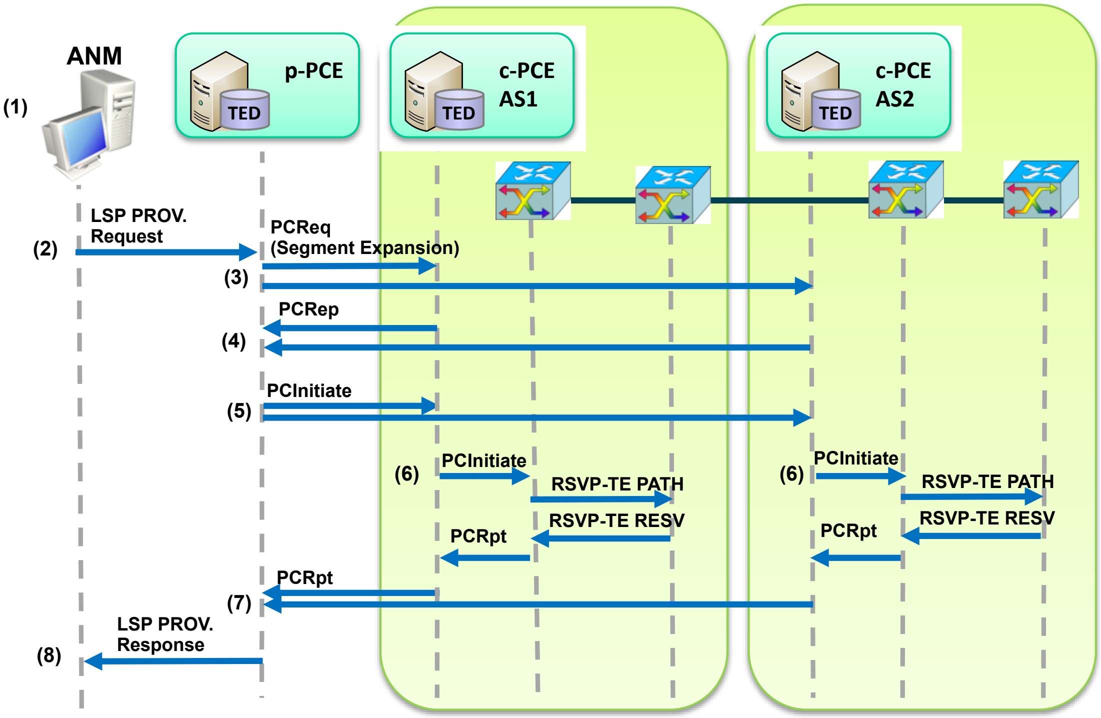


Figure 4. Message flow in architecture with Stateful H-PCE and per-child instantiation (SF-PD architecture)

1) SF-PD Procedure

Figure 4 will be used to explain the procedure in the SF-PD case. Following the SDN controller (ANM) request for a path establishment (2), the pPCE , using the local TED, calculates the domain sequence and then asks all the cPCEs for expanding the domain path within their respective domains (3). After receiving the replies from all the cPCE (4), the pPCE chooses a frequency slot satisfying the request constraints (in terms of slot width) that is free in all the domains and in the involved inter-domain links and performs Explicit Label Control (ELC) by adding a label subobiect after every hop in the received EROs. It's worth noting that, to perform correctly the spectrum assignment phase, the pPCE receives from the cPCEs the spectrum availability of the inter-domain links via the BGP-LS protocol and that of the different segments by means of dedicated extensions of the PCRep PCEP message used for segment expansion.

The EROs obtained after this step are then used inside the PCInitiate message sent to the respective cPCE (5). In turn, every cPCE forwards the PCInitiate message to the domain ingress node (6) that acts as the signaling source, starting a RSVP-TE session limited to the local domain. Upon completion of the signaling procedure, the PCRpt message, conveying the session's status, is sent back to the domain cPCE that, in turn, forwards it to the pPCE (7). At last, the pPCE notifies the ANM with the results of the procedure (8).

The LSP provisioning request could be based either on a management infterface (like Netconf or a REST API) or reuse the PCEP Initiate and Reports messages, as they convey the necessary information.

III. Protocol Extensions

The procedures described in the previous section require extensions in existing protocols.

A. PCEP Stateful Extensions

In order to implement PCE GMPLS, H-PCE and stateful capabilities and to drive proper operation in the distributed test-bed, a number of PCEP extensions have been implemented, tested and validated. .

1) Extensions in OPEN

To advertise novel path computation capabilities a number of TLVs have been enclosed in the OPEN object during the session handshake between cPCEs and pPCE. In particular the following TLVs have been implemented:

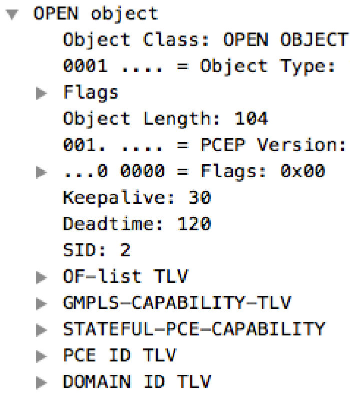
I) GMPLS Capability TLV (value: 14) [12], to advertise that the PCE is capable to perform path computation of paths with spectrum switching capability

II) Stateful PCE Capability TLV (value 16) [4] [5], to advertise that the PCE is stateful and if initiation capability is supported. In scenario SF-PD the initiation capability is supported by both pPCE and cPCEs.

III) PCE ID TLV (value 32769) [13], to advertise the unique identification of the PCE in the network. In this case, IPv4 address is considered.

IV) Domain ID TLV (value 32771) [13], to advertise the identification of the domain controlled by the PCE. in this case autonomous system identification is considered.

V) OF-Code list TLV (value 4) [14] to advertise the list of supported objective functions in the path computation algorithms implemented locally.



```
▼ OPEN object
    Object Class: OPEN OBJECT (1)
    0001 .... = Object Type: 1
  ▶ Flags
    Object Length: 104
    001. .... = PCEP Version: 1
  ▶ ...0 0000 = Flags: 0x00
    Keepalive: 30
    Deadtime: 120
    SID: 2
  ▶ OF-list TLV
  ▶ GMPLS-CAPABILITY-TLV
  ▶ STATEFUL-PCE-CAPABILITY
  ▶ PCE ID TLV
  ▶ DOMAIN ID TLV
```

Figure 5. OPEN with extended TLVs

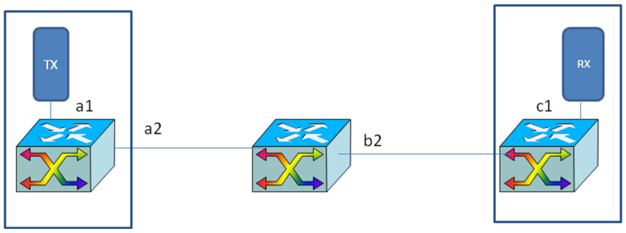


Figure 6. Interface reference for ERO format

2) LSP Concatenation (ERO format)

The Explicit Route Object is used in both PCEP to specify the details of the path in a path computation reply message and in an instantiation. The ERO format is as follows:

- By default, ERO unnumbered interfaces represent outgoing TE links (both intra-domain and inter-domain TE links)
- To convey information about the ingress port, if the first node router ID (router IDs and node IDs are assumed to be equal) appears twice consecutively in the ERO, it is assumed that the first subobject specifies the ingress interface (facing the client).
- The ERO may end with a IPv4 prefix address (/32) meaning that the LSP ends at the egress node incoming interface or an unnumbered interface, meaning that the LSP ends at the output interface with an additional cross-connect.
- In all the cases, explicit label control (ELC) conveys information about the outgoing label (frequency slot) that will be used by the downstream node in switching. The label specifies both the slot center frequency (n) and the slot width (m).

In the case the endpoint transponders are not indicated in the ANM→pPCE PCReq, the pPCE has to select the transponders and enclose their indication within the ERO of the final PCRep (SL-E2E case) or of the PCInit sent to the cPCEs of the ingress and egress domains, as well as in the PCRpt sent to the ANM (SF-PD case). In order to enable the selection of the end-points transponders, the ERO provided by the pPCE to the source cPCE has to be expanded accordingly. To identify the selected transponders within the ERO without introducing a new ERO subobject, the following agreement is proposed:

- The first and the second ERO subobjects of the e2e path are referred to the selected transponder located at the source node. The penultimate and the ultimate ERO subobjects are referred to the selected transponder at the destination.
- The first (and the penultimate) ERO subobject is of type "Unnumbered" and identifies the transponder.

The second (and the last) ERO subobject is of type "Label" and identifies the m and n values "generated" by the transponder (source node case) or used to switch the traffic to the receiver and to tune its central frequency (destination node case). The value of m and n may in general be different from those associated to the outgoing optical link. In particular m can be different based on the specific architecture of the optical node.

Referring to Figure 6, the ERO is: A[a1], label(tx), A[a2], label, B[b2], label, C[c1], label (rx).

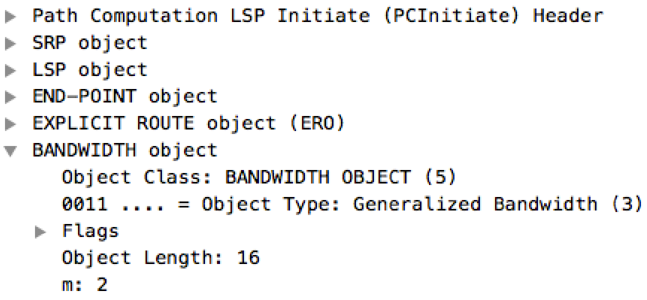
In particular label(tx) is the information needed by the transponder to setup the tunable laser and the electrical (or optical) filter. In general, such values could be different from the label values of the outgoing link interface a2. This because the granularity required from such devices could be different from the standard flexi-grid value (6.25GHz).

Figure 8 contains a Wireshark capture with the detail of the ERO of a segment in one domain.

3) PCEP Initiate

In the SF-PD architecture the PC Initiate (PCInit) and PC Report (PCRpt) messages are considered to trigger the instantiation of each segment.

The PCInit message specifies that a certain LSP has to be instantiated or removed in the network. The PCInit message includes the endpoints and the ERO. Moreover, it includes the LSP object, specifying the identification of the LSP (including the symbolic LSP name). In case of novel instantiation, such identification is set to a default value since a unique id (the P_LSP id) will be provided either by the cPCE or the controlled network once the LSP has been configured in the data plane and stored in the local stateful database. In case of LSP removal request, the LSP id is provided (in the P_LSP id field of the LSP object). A PCInit message is sent also in the case a LSP is requested to be both computed and instantiated. In this case the ERO (which is mandatory object in the PCInit) encloses only the endpoints. The match between endpoints and ERO subojects triggers possible path computation at the receiver PCE. Such option is utilized by ABNO when it triggers a new instantiation towards pPCE.



```
▶ Path Computation LSP Initiate (PCInitiate) Header
▶ SRP object
▶ LSP object
▶ END-POINT object
▶ EXPLICIT ROUTE object (ERO)
▼ BANDWIDTH object
    Object Class: BANDWIDTH OBJECT (5)
    0011 .... = Object Type: Generalized Bandwidth (3)
  ▶ Flags
    Object Length: 16
    m: 2
```

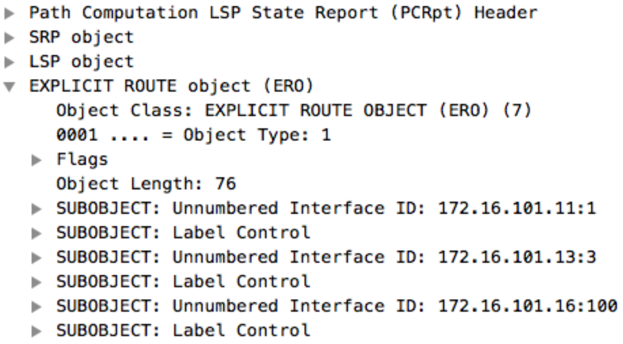
Figure 7. PCEP Initiate and Generalized Bandwidth Object

4) Requested Spectrum (Bandwidth)

The PCEP Initiate, as well as the PCEP Request, needs to include the quantity of spectrum needed. The requested spectrum is included in the Generalized Bandwidth Object [12], in particular encoded as SSON traffic parameters. The parameter included is the m, which multiplied by 6,25 GHz represents the requested spectrum width. A Wireshark capture of the described Generalized Bandwidth Object is shown in Figure 7.

5) PCEP Report

The PC Report (PCRpt) message is sent to advertise the current status of an LSP upon initiation, modification events or for LSP-DB synchronization. Besides SRP object, the PCRpt encloses the current LSP operational status within the LSP object (i.e., LSP up, LSP active, LSP going down, LSP removed indicated in the Operational bit flag), also including the LSP identifiers TLV (in this case IPv4 identifiers have been used) in terms of RSVP-TE session (LSP id and TUNNEL id of the established paths) and LSP symbolic path name. In particular the Delegate flag is set by cPCEs to 1, indicating that the cPCE delegates path computation to pPCE. The PCRpt generation trigger is left to the different PCE implementations.



```
▶ Path Computation LSP State Report (PCRpt) Header
▶ SRP object
▶ LSP object
▼ EXPLICIT ROUTE object (ERO)
    Object Class: EXPLICIT ROUTE OBJECT (ERO) (7)
    0001 .... = Object Type: 1
  ▶ Flags
    Object Length: 76
  ▶ SUBOBJECT: Unnumbered Interface ID: 172.16.101.11:1
  ▶ SUBOBJECT: Label Control
  ▶ SUBOBJECT: Unnumbered Interface ID: 172.16.101.13:3
  ▶ SUBOBJECT: Label Control
  ▶ SUBOBJECT: Unnumbered Interface ID: 172.16.101.16:100
  ▶ SUBOBJECT: Label Control
```

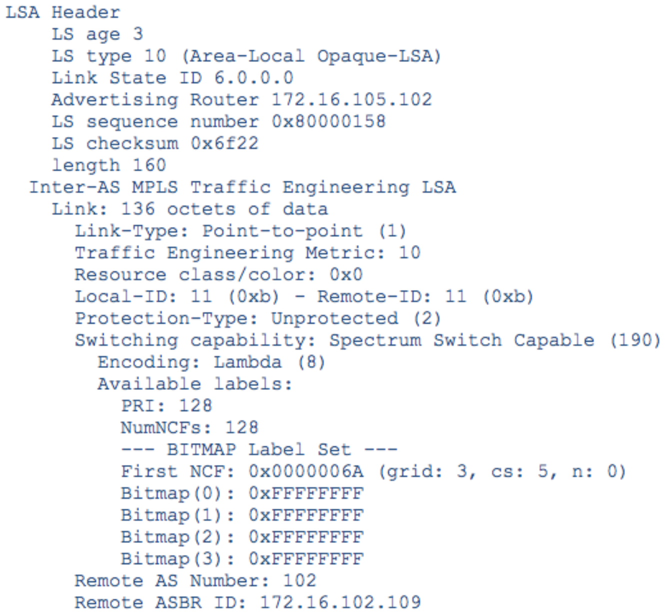
Figure 8 PCEP Report & ERO Example

B. OSPF-TE Extensions

The OSFP-TE protocol has been extended to support flexi-grid networks. The extensions inherit the previous work done in the scope of Wavelength Switched Optical Networks (WSON), for which the framework was defined in [15]. With the exception of wavelength-specific availability information, the connectivity topology and node capabilities are the same, which can be advertised by the GMPLS routing protocol.

For Elastic optical networks based on flexi-grid, a set of non-overlapping available frequency ranges should be disseminated in order to allow efficient resource management of flexi-grid DWDM links and RSA procedures, i.e., in the flexi-grid case, the available frequency ranges are advertised for the link instead of the specific "wavelengths".

The proposed extensions, being pushed for standardization in [10], mainly disseminate the status of the Nominal Central Frequencies. Such extensions are carried into the Interface Switching Capability Descriptor (ISCD), and more specifically in the Switching Capability Specific Information (SCSI).



```
LSA Header
    LS age 3
    LS type 10 (Area-Local Opaque-LSA)
    Link State ID 6.0.0.0
    Advertising Router 172.16.105.102
    LS sequence number 0x80000158
    LS checksum 0x6f22
    length 160
  Inter-AS MPLS Traffic Engineering LSA
    Link: 136 octets of data
      Link-Type: Point-to-point (1)
      Traffic Engineering Metric: 10
      Resource class/color: 0x0
      Local-ID: 11 (0xb) - Remote-ID: 11 (0xb)
      Protection-Type: Unprotected (2)
      Switching capability: Spectrum Switch Capable (190)
        Encoding: Lambda (8)
        Available labels:
          PRI: 128
          NumNCFs: 128
          --- BITMAP Label Set ---
          First NCF: 0x0000006A (grid: 3, cs: 5, n: 0)
          Bitmap(0): 0xFFFFFFFF
          Bitmap(1): 0xFFFFFFFF
          Bitmap(2): 0xFFFFFFFF
          Bitmap(3): 0xFFFFFFFF
      Remote AS Number: 102
      Remote ASBR ID: 172.16.102.109
```

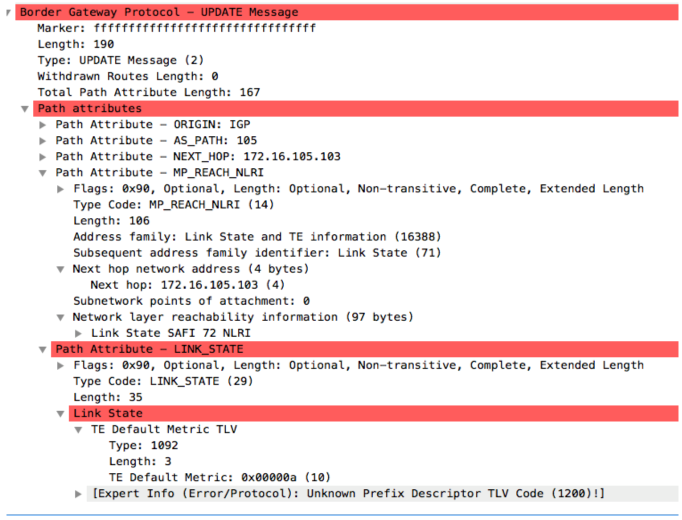
Figure 9 OSPF-TE Inter-AS log

C. BGP-LS Extensions

The BGP-4 protocol has been extended by IETF to support the exchange of link-state information between two entities [7]. In the context of EON, and in particular, multi-domain optical networks, BGP-LS can be used as a mean to send TE information (or, in case of limited number of domains and network nodes, the entire TE database) to a PCE. In the described multi-domain scenario, cPCE exports TED to pPCE by means of BGP-LS, including nodes and TE links description.

After the BGP-LS session has been established, in order to export the topology, a peer can send UPDATE messages including the MP_REACH attribute. How often are the updates sent later depend on the specific implementation of each cPCE. The Network Layer Reachability Information (NLRI) contains the information of nodes and links. The node is mainly described in terms of IPv4 router ID and AS, while a EON link needs to be described by its source and destination (node, interface and domain it belongs) and the TE information. Optical Node and Link NLRI needs to indicate the source of the information in the protocol ID (3 for OSPF sources, 5 for static configuration). In the experimental test-bed, As in the node case, the identifiers are set to L1 Optical Topology value. The source and destination endpoints of the link are indicated by IPv4 addresses of the nodes, which are encoded in the Link NLRI, in the Local Node Descriptors TLV (i.e., information of the source) and Remote Node Descriptors TLV (i.e., information of the destination), in the IGP Router-ID fields. In addition, the endpoints of the link are characterized by the Autonomous System ID (in the Autonomous System TLV) and Area ID. Thus, in the case of a link between two Autonomous System there will be different AS IDs in the Local Node Descriptor and Remote Node Descriptor. In the case of intra-domain link both IDs will be the same.

In a EON link it is necessary to indicate not only the source and destination nodes, but also the interface where the fiber is connected. The interfaces are identified by the use of unnumbered interfaces, which are encoded in the link local/remote identifiers TLV of the Link NLRI.



```
▾ Border Gateway Protocol - UPDATE Message
    Marker: ffffffffffffffffffffffffffffffff
    Length: 190
    Type: UPDATE Message (2)
    Withdrawn Routes Length: 0
    Total Path Attribute Length: 167
  ▾ Path attributes
    ▸ Path Attribute - ORIGIN: IGP
    ▸ Path Attribute - AS_PATH: 105
    ▸ Path Attribute - NEXT_HOP: 172.16.105.103
    ▾ Path Attribute - MP_REACH_NLRI
      ▸ Flags: 0x90, Optional, Length: Optional, Non-transitive, Complete, Extended Length
        Type Code: MP_REACH_NLRI (14)
        Length: 106
        Address family: Link State and TE information (16388)
        Subsequent address family identifier: Link State (71)
      ▾ Next hop network address (4 bytes)
          Next hop: 172.16.105.103 (4)
        Subnetwork points of attachment: 0
      ▾ Network layer reachability information (97 bytes)
        ▸ Link State SAFI 72 NLRI
    ▾ Path Attribute - LINK_STATE
      ▸ Flags: 0x90, Optional, Length: Optional, Non-transitive, Complete, Extended Length
        Type Code: LINK_STATE (29)
        Length: 35
      ▾ Link State
        ▾ TE Default Metric TLV
            Type: 1092
            Length: 3
            TE Default Metric: 0x00000a (10)
        ▸ [Expert Info (Error/Protocol): Unknown Prefix Descriptor TLV Code (1200)!]
```

Figure 10. BGP-LS Update message

The TE information of the EON link attributes is carried in the BGP-LS Attribute. The exchanged information includes the IPv4 router IDs of both local and remote nodes, the maximum, unreserved, reservable bandwidth, the TE Default Metric (type code 1092), the SRLG (type code 1096). With respect to the IETF draft, a novel proposed parameter has been included for EON enclosing the available labels expressed as the bitmap of available/occupied nominal central frequencies (proposed type code 1200). In this way, complete domain TE, including detailed per-link spectrum slot occupancy info collected by local cPCE can be exported towards the pPCE, thus enabling end-to-end path/domain computation including advanced spectrum suggestion/assignment.

D. RSVP-TE Extensions

New extensions have been identified for the RSVP-TE signaling protocol; some of them are shared with other involved protocols (PCEP, BGP-LS, OSPF-TE). In particular, a new label format based on 64 bit encoding of the central frequency and slot width [16] and the ability to disseminate the status of the nominal central frequencies on a per link basis, using a bitmap format encoding within the LABEL_SET object have been implemented. A new sender traffic parameters or SENDER_TSPEC (included in the sender descriptor of the Path message) as well as a new FLOWSPEC object in the flow descriptor (Resv message) having new types to convey the desired and assigned frequency slot width, respectively, have been defined [11]. At last, a new switching capability value (190) has been proposed to indicate the SSON media layer.

IV. Experimental Validation

A. Idealist Multi-partner Test-bed

The Idealist Multi-partner Control Plane Test-bed interconnects four European research institutions, located in Madrid (Telefónica I+D), Barcelona (CTTC), Torino (TI) and Pisa (CNIT). The test-bed physical topology is depicted in Figure 11. Partners' premises are connected (at the control plane level) by means of dedicated IPsec tunnels. The resulting low level connectivity layout is a hub, centered at CTTC. Static routing entries provide full connectivity between partners' private addresses, secured and isolated from the rest of Internet traffic. On top of this distributed control plane connectivity network, logical relationships between PCEs are established, in particular between Telefónica I+D PCE, acting as pPCE, and the other PCEs, acting as cPCE, as shown in Figure 11. The PCEs of the test-bed have been independently developed by each partner. BGP-LS speakers are implanted by each partner [17].

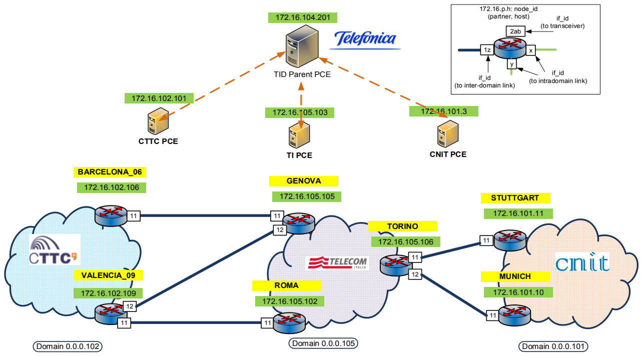
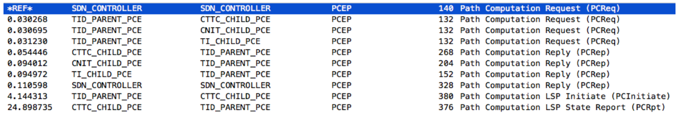


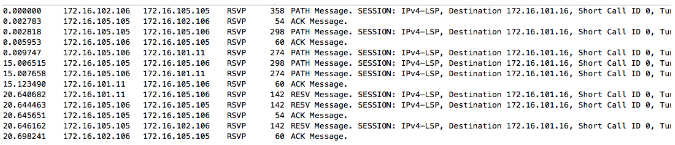
Figure 11. Topology of Muti-partner Idealist Test-bed

Telefónica I+D pPCE is an open source multi-threaded application developed in Java 1.6 [18]. It accepts sessions from cPCEs, maintaining each session with a specific thread which handles all the messages exchange. Also, a dedicated thread is used for each BGP-LS session, building the multi-domain TE Database (TED), in which the nodes are domains, and the edges are the inter-domain links, and the reachability information is obtained by node advertisements. CNIT test-bed comprises 7 C++-based EON controllers capable of dynamically configuring co-located SSS by means of USB interface and a C++ based cPCE performing advanced impairment-aware computation [19][20][21]. CTTC domain emulates a 14 node mesh network that represents a national (Spanish) photonic transport segment and a stateful cPCE. TI test-bed is composed of 6 Linux boxes running GMPLS-based control plane processes, whose architecture is in line with Rec. ITU-T G.8080 in terms of architectural components, emulating a single network node each.



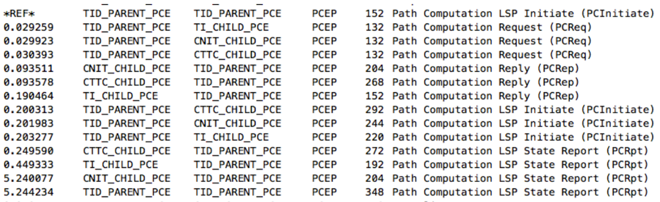
```
*REF*       SDN_CONTROLLER    SDN_CONTROLLER    PCEP         140 Path Computation Request (PCReq)
0.030268    TID_PARENT_PCE    CTTC_CHILD_PCE    PCEP         132 Path Computation Request (PCReq)
0.030695    TID_PARENT_PCE    CNIT_CHILD_PCE    PCEP         132 Path Computation Request (PCReq)
0.031230    TID_PARENT_PCE    TI_CHILD_PCE      PCEP         132 Path Computation Request (PCReq)
0.054446    CTTC_CHILD_PCE    TID_PARENT_PCE    PCEP         268 Path Computation Reply (PCRep)
0.094012    CNIT_CHILD_PCE    TID_PARENT_PCE    PCEP         204 Path Computation Reply (PCRep)
0.094972    TI_CHILD_PCE      TID_PARENT_PCE    PCEP         152 Path Computation Reply (PCRep)
0.110598    SDN_CONTROLLER    SDN_CONTROLLER    PCEP         328 Path Computation Reply (PCRep)
4.144313    TID_PARENT_PCE    CTTC_CHILD_PCE    PCEP         380 Path Computation LSP Initiate (PCInitiate)
24.898735   CTTC_CHILD_PCE    TID_PARENT_PCE    PCEP         376 Path Computation LSP State Report (PCRpt)
```

Figure 12. Capture of PCEP message flow in SL-E2E



```
0.000000    172.16.102.106  172.16.105.105  RSVP   358 PATH Message. SESSION: IPv4-LSP, Destination 172.16.101.16, Short Call ID 0, Tu
0.002783    172.16.105.105  172.16.102.106  RSVP    54 ACK Message.
0.002818    172.16.105.105  172.16.105.106  RSVP   298 PATH Message. SESSION: IPv4-LSP, Destination 172.16.101.16, Short Call ID 0, Tu
0.005953    172.16.105.106  172.16.105.105  RSVP    60 ACK Message.
0.009747    172.16.105.106  172.16.101.11   RSVP   274 PATH Message. SESSION: IPv4-LSP, Destination 172.16.101.16, Short Call ID 0, Tu
15.006515   172.16.105.105  172.16.105.106  RSVP   298 PATH Message. SESSION: IPv4-LSP, Destination 172.16.101.16, Short Call ID 0, Tu
15.007658   172.16.105.106  172.16.101.11   RSVP   274 PATH Message. SESSION: IPv4-LSP, Destination 172.16.101.16, Short Call ID 0, Tu
15.123490   172.16.101.11   172.16.105.106  RSVP    60 ACK Message.
20.640682   172.16.101.11   172.16.105.106  RSVP   142 RESV Message. SESSION: IPv4-LSP, Destination 172.16.101.16, Short Call ID 0, Tu
20.644463   172.16.105.106  172.16.105.105  RSVP   142 RESV Message. SESSION: IPv4-LSP, Destination 172.16.101.16, Short Call ID 0, Tu
20.645651   172.16.105.105  172.16.105.106  RSVP    54 ACK Message.
20.646162   172.16.105.105  172.16.102.106  RSVP   142 RESV Message. SESSION: IPv4-LSP, Destination 172.16.101.16, Short Call ID 0, Tu
20.698241   172.16.102.106  172.16.105.105  RSVP    60 ACK Message.
```

Figure 13. RSVP-TE capture at Telecom Italia border node "Torino".



```
*REF*      TID_PARENT_PCE   TID_PARENT_PCE   PCEP   152 Path Computation LSP Initiate (PCInitiate)
0.029259   TID_PARENT_PCE   TI_CHILD_PCE     PCEP   132 Path Computation Request (PCReq)
0.029923   TID_PARENT_PCE   CNIT_CHILD_PCE   PCEP   132 Path Computation Request (PCReq)
0.030393   TID_PARENT_PCE   CTTC_CHILD_PCE   PCEP   132 Path Computation Request (PCReq)
0.093511   CNIT_CHILD_PCE   TID_PARENT_PCE   PCEP   204 Path Computation Reply (PCRep)
0.093578   CTTC_CHILD_PCE   TID_PARENT_PCE   PCEP   268 Path Computation Reply (PCRep)
0.190464   TI_CHILD_PCE     TID_PARENT_PCE   PCEP   152 Path Computation Reply (PCRep)
0.200313   TID_PARENT_PCE   CTTC_CHILD_PCE   PCEP   292 Path Computation LSP Initiate (PCInitiate)
0.201983   TID_PARENT_PCE   CNIT_CHILD_PCE   PCEP   244 Path Computation LSP Initiate (PCInitiate)
0.203277   TID_PARENT_PCE   TI_CHILD_PCE     PCEP   220 Path Computation LSP Initiate (PCInitiate)
0.249590   CTTC_CHILD_PCE   TID_PARENT_PCE   PCEP   272 Path Computation LSP State Report (PCRpt)
0.449333   TI_CHILD_PCE     TID_PARENT_PCE   PCEP   192 Path Computation LSP State Report (PCRpt)
5.240077   CNIT_CHILD_PCE   TID_PARENT_PCE   PCEP   204 Path Computation LSP State Report (PCRpt)
5.244234   TID_PARENT_PCE   TID_PARENT_PCE   PCEP   348 Path Computation LSP State Report (PCRpt)
```

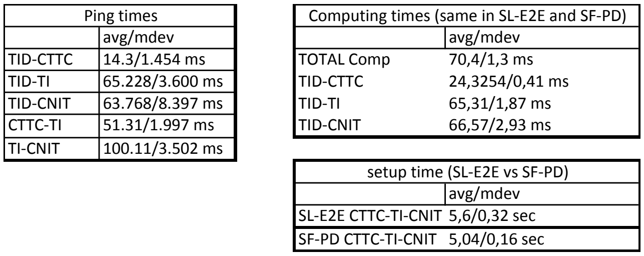
Figure 14 Capture of message flow in SF-PD

B. Performance Evaluation and analysis of SL-E2E

First of all, the architecture has been fully validated functionally, and the interoperability of four PCEP implementations has been achieved. The details of the protocol interoperability have been reported to IETF [17]. Figure 12 shows a Wireshark capture in the parent PCE machine, showing the path computation interactions and the later initiation procedure when architecture SL-E2E is used. In the example, a media channel with m = 2 is requested from node 172.16.102.101 interface 1000, in CTTC domain, to node 172.16.101.16 interface 100 in CNIT domain. Figure 13 shows the RSVP-TE interactions that are triggered by the PCEP instantiation message. The experiment shows that the total computation time, including all the interactions between parent and child PCEs is 68 ms. In the example, three domains are involved in the computation and instantiation. The initiation time is 5.7 seconds. Most of the time of the initiation is due to the configuration of the real OXC WSS based flexi-grid nodes in CNIT test-bed. The total initiation time is shown in Eq. 1, which includes the time to reach the parent PCE (in both ways, $RTT_{ANM-PPCE}$), the computing time of the parent PCE (T_{comp_PPCE}), the maximum of the queries to the child PCEs ($T_{PPCE-X-CPCE}$, where X is CNIT, CTTC or TI, and includes the time to reach the child PCE and the computing time of the child PCE), the time to query (and get and answer) the head end node ($RTT_{ANM-HEAD-NODE}$), and the end-to-end signaling time (sum of the RSVP times per domain and the RTT of the inter-domain links). In this architecture, the computing time is dominated by the slowest response time of a child PCE, as the requests are in parallel, but the setup time increases with the number of domain, as there is an end-to-end signaling session.

$$T_{ini-SL-E2E} = RTT_{ANM-PPCE} + T_{COMP\_PPCE}$$
$$\max\left(T_{PPCE-CNIT-CPCE}, T_{PPCE-CTTC-CPCE}, T_{PPCE-TI-CPCE}\right)$$
$$+RTT_{ANM-HEAD-NODE} + \sum_{each\_domain} T_{RSVP} + \sum_{each\_id\_link} RTT_i$$

Eq. 1 Total Initiation time in SL-E2E



Ping times	
	avg/mdev
TID-CTTC	14.3/1.454 ms
TID-TI	65.228/3.600 ms
TID-CNIT	63.768/8.397 ms
CTTC-TI	51.31/1.997 ms
TI-CNIT	100.11/3.502 ms

Computing times (same in SL-E2E and SF-PD)	
	avg/mdev
TOTAL Comp	70,4/1,3 ms
TID-CTTC	24,3254/0,41 ms
TID-TI	65,31/1,87 ms
TID-CNIT	66,57/2,93 ms

setup time (SL-E2E vs SF-PD)	
	avg/mdev
SL-E2E CTTC-TI-CNIT	5,6/0,32 sec
SF-PD CTTC-TI-CNIT	5,04/0,16 sec

Figure 15. Experimental Results

It has to be taken into account that both CTTC and TI test-beds run on emulated nodes, and thus set-up time is faster. The latencies between the different components is shown in, that shows the mean times of the important steps as well.

C. Performance Evaluation and analysis of SF-PD

Figure 14 shows the message flow (Wireshark capture) in the parent PCE machine, including the path computation interactions and the later initiation procedure using the SF-PD architecture. The same end-points as in the previous example are used. In this experiment, the total computation time, including all the interactions between parent and child PCEs is similar to the SL-E2E case, as the computation procedure remains unchanged. However, the total initiation time, 5.25 seconds, is lower than in the SL-E2E case, as the per-domain initiations are performed in parallel. The total initiation time is shown in Eq. 2, which includes the time to reach the parent PCE (in both ways, $RTT_{ANM-PPCE}$), the computing time of the parent PCE (T_{comp_PPCE}), the maximum of the queries to the child PCEs ($T_{PPCE-X-CPCE}$, where X is CNIT, CTTC or TI, and includes the time to reach the child PCE and the computing time of the child PCE), the maximum of the times to provision in each domain (summing the round trip time from parent PCE to the child PCE, and the signaling time in the domain). As there is no end-to-end signaling, no messages are exchanged between domains. The time of the initiation in this case is dominated by the longer setup time, which corresponds to the real OXC WSS based flexi-grid nodes in CNIT test-bed.

$$T_{ini-SF-PD} = RTT_{ANM-PPCE} + T_{COMP\_PPCE}$$
$$\max\left(T_{PPCE-CNIT-CPCE}, T_{PPCE-CTTC-CPCE}, T_{PPCE-TI-CPCE}\right)$$
$$+\max\begin{pmatrix} \left(RTT_{PPCE-CTTC} + T_{RSVP-CTTC}\right), \\ \left(RTT_{PPCE-TI} + T_{RSVP-TI}\right), \\ \left(RTT_{PPCE-CNIT} + T_{RSVP-CNIT}\right) \end{pmatrix}$$

Eq. 2 Total Initiation time in SF-PD

Both architectures are suitable for the control of multi-domain elastic optical networks. Even though the parallelism of SF-PD gives an advantage in terms of performance, it adds the complexity of maintaining the end-to-end LSPs in the parent PCE.

V. Conclusions

In this work an extended demonstration of multi-domain EON control plane based on HPCE architecture was implemented, evaluated and validated in the scope of the IDEALIST European Project. High degree of interoperability achieved in the multi-partner multi-platform distributed control plane EON test-bed was achieved by employing the most recent PCEP, OSPF-TE, BGP-LS and RSVP-TE protocol extensions. Two architectural scenarios, based on stateless PCE performing end-to-end instantiation and stateful PCE performing per-domain instantiation and stitching were implemented and fully tested, respectively. Experimental results showed path computation, including all protocol interactions, path computation and LSP setup times. Moreover, extended control plane messages and operation for each scenario were detailed, highlighting the extensions conceived for EON and multi-domain scenario. The adoption of BGP-LS extensions fully enabled multi-domain TE and was demonstrated in a limited number of domains. The results provided in this work aim at representing the current state-of-the-art of the research in EON control plane and the reference benchmark for future research activities and extensions in the context of multi domain optical networks.

Acknowledgment

This work was supported by the FP-7 IDEALIST project under grant agreement number 317999. This paper is an extended version of the work in [22].

References

[1] M. Bohn et al.," Elastic Optical Networks: the Vision of the ICT Project IDEALIST", FuNeMS 2013.

[2] http://www.ict-idealist.eu

[3] D. King and A. Farrel. “A PCE-based Architecture for Application-based Network Operations”, IETF draft-farrkingel-pce-abno-architecture-13

[4] E. Crabbe et al., “PCEP Extensions for Stateful PCE” IETFdraft-ietf-pce-stateful-pce-11

[5] E. Crabbe et al., “PCEP extensions for PCE-initiated LSP setup in a stateful PCE model,” IETF draft-ietf-pce-pce-initiated-lsp-04

[6] F. Paolucci et al., "Experimenting Hierarchical PCE architecture in a distributed multi-platform control plane testbed", OM3G.3, OFC 2012.

[7] H.Gredler et al., “North-Bound Distribution of Link-State and TE Information using BGP”, IETF draft-ietf-idr-ls-distribution-06

[8] M. Cuaresma et al. “Experimental Demonstration of H-PCE with BPG-LS in elastic optical networks”, ECOC 2013

[9] D. King and A. Farrel, “The Application of the Path Computation Element Architecture to the Determination of a Sequence of Domains in MPLS and GMPLS”, RFC 6805

[10] F. Zhang et al. “GMPLS OSPF-TE Extensions in support of Flexible Grid in DWDM Networks”, IETF draft-zhang-ccamp-flexible-grid-ospf-ext-04.

[11] F. Zhang et al. “RSVP-TE Signaling Extensions in support of Flexible Grid”, IETF draft-ietf-ccamp-flexible-grid-rsvp-te-ext-00.

[12] C. Margaria et al., “PCEP extensions for GMPLS”, IETF draft-ietf-pce-gmpls-pcep-extensions-10.

[13] F. Zhang et al., “Extensions to Path Computation Element Communication Protocol (PCEP) for Hierarchical Path Computation Elements (PCE)”, IETF draft-ietf-pce-hierarchy-extensions-02.

[14] JL. Le Roux,et al., “Encoding of Objective Functions in the Path Computation Element Communication Protocol (PCEP)”, RFC 5541.

[15] Y. Lee et al., “Framework for GMPLS and Path Computation Element (PCE) Control of Wavelength Switched Optical Networks (WSONs)”, RFC 6163.

[16] A. Farrel et al., “Generalized Labels for the Flexi-Grid in Lambda Switch Capable (LSC) Label Switching Routers”, IETF draft-ietf-ccamp-flexigrid-lambda-label-03.

[17] H. Gredler et al., "BGP Link-State Information Distribution Implementation Report", IETF draft-ietf-idr-ls-distribution-impl-04.

[18] Netphony Open Source PCE https://github.com/telefonicaid/netphony-pce

[19] F. Paolucci, A. Castro, F. Fresi, M. Imran, A. Giorgetti, B. B. Bhownik, G. Berrettini, G. Meloni, F. Cugini, L. Velasco, L. Potì, P. Castoldi, Active PCE demonstration performing elastic operations and hitless defragmentation in flexible grid optical networks , Photonic Network Communications, Springer, vol. 29, issue 1, 2014, pp 57-66

[20] N. Sambo, F. Paolucci, G. Meloni, F. Fresi, L. Potì, P. Castoldi, Control of Frequency Conversion and Defragmentation for Super-Channels [Invited], IEEE/OSA Journal of Optical Communications and Networking, vol. 7, n. 1, 2015, pp A126-A134

[21] F. Cugini, F. Fresi, F. Paolucci, G. Meloni, N. Sambo, A. Giorgetti, T. Foggi, L. Potí, P. Castoldi, Active Stateful PCE With Hitless LDPC Code Adaptation [Invited], IEEE/OSA Journal of Optical Communications and Networking, vol. 7, n. 2, 2015, pp A268-A276

[22] O.G. de Dios, R. Casellas, R. Morro, F. Paolucci, V. Lopez, R. Martinez, R. Munoz, R. Vilalta, P. Castoldi, First Multi-partner Demonstration of BGP-LS enabled Inter-domain EON control with H-PCE, Optical Fiber Conference 2015 Tech. Dig., paper Th1A.4, OSA, March 2015.